caGEDA

UPCI

http://bioinformatics.upmc.edu/GE2/GEDA.html

Why caGEDA?

- Evaluative comparisons of new methods of analysis is rarely conducted - and is needed
- Normalization methods are not well understood
- Performance characteristics of tests for identifying differentially expressed genes are understudied
- Optimal combinations of normalization -> feature selection -> sample classification algorithms have not yet been determined
- caGEDA was/is designed with cancer researchers in mind

caGEDA

'Why Not Just Use…'

- BioConductor
- MeV/TM4
- OncoMine
- BRBArray Tools
- GEDP***
- Others…
- Commercial software

*** microarray data repository – core to caBIG!!!

•Please do! Some very nice options!

•Some require downloads/registration
•Some require programming
•Some are not open source

•Every new microarray data set is another opportunity to identify generally optimized methods of analysis

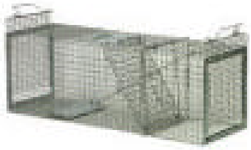
•Training and adoption is efficient with a web application

Normalization	Concept	References
Reference Gene/ Sample Subset Methods		
Housekeeping Genes	Selection of a set of genes as controls; each value in an array is normalized using the mean of this subset	Lee et al., 2001; Vandesompele et.al., 2002
'Globalization' Method	Each value in an array is normalized using the global mean of all arrays	Velculescu et al., 1999
Loess 1: Normalization by self- consistency and local regression	Normalize pairs or groups of arrays relative to each other by iteratively maximizing the consistency of relative expression levels among them. Genes are consistent if their relative expression values do not change after global normalization. The original data are normalized using the consistent set and local regression	Kepler et al., 2002
Iterative Invariant Set Normalization	Find gene set with unchanged ranks in expression in both groups; use an iterative procedure to identify invariant set as those probes with proportion rank difference (PRD) < 0.003 (low rank) or < 0.007 (high rank genes)	Li & Wong, 2002
Microarray Sample Pool	Normalize all samples using an ensemble sample (MSP) as the reference array	Yang YH et al., 2002
Statistical Methods		
Variance Stabilization	Normalization by the arsinh function h(y) = *g* arsinh(*a*+*by*) with model parameters *a* and *b* estimated by likelihood	Huber et al, 2002
Variance Stabilization	Stabilizes asymptotic variance over the full range of expression intensity. Finds a transformation for a regression model such that the variance is constant over the range of the dependent variable	Durbin et al., 2002
Dye Channel Control Spot Scaling	Expression values normalized by scaling cy5 values so that mean cy5 & cy3 values in control spots are same	Cavalieri et al., 2000
Loess 2:Local mean normalization	Calculation of local mean (using regression) and distance of this mean from each ratio is the corrected ratio. Results in mean intensity ratio of 1	Colantuoni et al., 2002
Loess 3:Local variance correction	Expression ratios made to have same local standard deviation calculated by loess and the intensity is represented as a Z-score	Colantuoni et al., 2002
Loess 4: Loess Local Regression	Intensity-dependent normalization achieved using the lowess function c(A), specifically log(R/G)corr= log(R/G)-c(A)	Yang YH et al, 2002
Log inverse ratio global normalization	Shift the log ratios by correction factor log(R/G)corr= log(R/G)-c where c = log(G/R); center of distribution shifted to 0	Yang YH et al., 2002
Variance regularization	Normalization factor is calculated using sum of both intensities, which is used to adjust the expression data in its log form	Quackenbush, 2002
Signal-Dependent Normalization	Center the mean of Cy3 & Cy5 log-ratio distributions	Workman et.al., 2002
Qspline	Quantiles from target and probe signals used to fit a smoothing B-spline	Workman et.al., 2002
Spot- Specific Normalization		
Adjustment for slide-specific effect	Ratio-based adjustments: normalize using error factor from simulations; categorical adjustments: use Bartlett's method	Tsodikov et al., 2002
Spatial Normalization	Subtract local signal estimates from log intensities or log ratios	Workman et.al., 2002

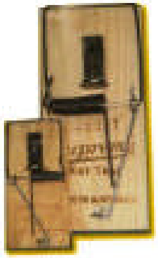
Test	Reference(s)
adaptive sign test	Boer et al., 2001
ANOVA	Kerr et al., 2000; Luo et al., 2002
BSS/WSS	Dudoit, 2002
diagnostic metric	Welsh et al., 2001
discriminative weighting	Bittner et al., 2000
empirical Bayes method	Newton et al. 2001
ideal discriminator method	Troyanskaya et al., 2002
local Bayesian Error test	Baldi and Long, 2001
log-odds tests	Lonnstedt and Speed, 2002
neighborhood analysis	Golub et al., 1999
nonparametric t-test	Garber et al., 2001;Troyanskaya et al., 2002
perfect discriminator permutation	Park et al. 2001
Pitman’s test	Herwig et al., 2001
ANOVA with bootstrap variance est.	Black & Doerge, 2002
significance analysis of microarrays (SAM)	Tusher et al., 2001
singular value decomposition	Alter et al., 2000; Wall et al. 2001;Ghosh, 2002
genetic algorithm	Li et al., 2001
partial least squares	Nguyen and Rocke, 2002
Welch test	Herwig et al., 2001
Z-ratio score	Quakenbush, 2002

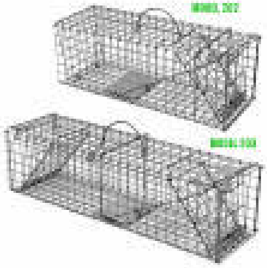
Algorithm	Reference(s)
BTSVQ	Sultan et al., 2002
cluster affinity search technique(CAST)	Ben-Dor et al., 1999
decision tree classification	Quinlan, 1996
deterministic annealing	Alon et al., 1999
gene shaving	Hastie et al., 2000
hierarchical clustering (various distances)	
k-means clustering	Eisen et al., 1998
Kohonen-clustering	Kohonen, 1982
logistic discrimination	Nguyen and Rocke, 2002
multidimensional scaling	Bittner et al., 2000
normalized cuts	Shi and Malik
neighbor joining	Saitou and Nei, 1987
nearest neighbor	Li et al., 2001; Theilhaber et al.2002
partitioning around medoids	Bozinov and Rahnenfuhrer, 2002
principle components analysis	(e.g., Luo et al., 2002)
quadratic discriminant analysis	Nguyen and Rocke, 2002
self-organizing maps	Dougherty et al., 2002
weighted voting	Golub et al., 1999;Yeang et al., 2001
Pitt-N Neighbors clustering	Lyons-Weiler et al., 2003

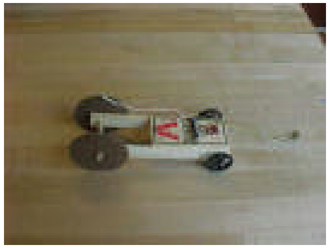
•

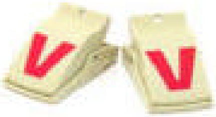


Too many methods.





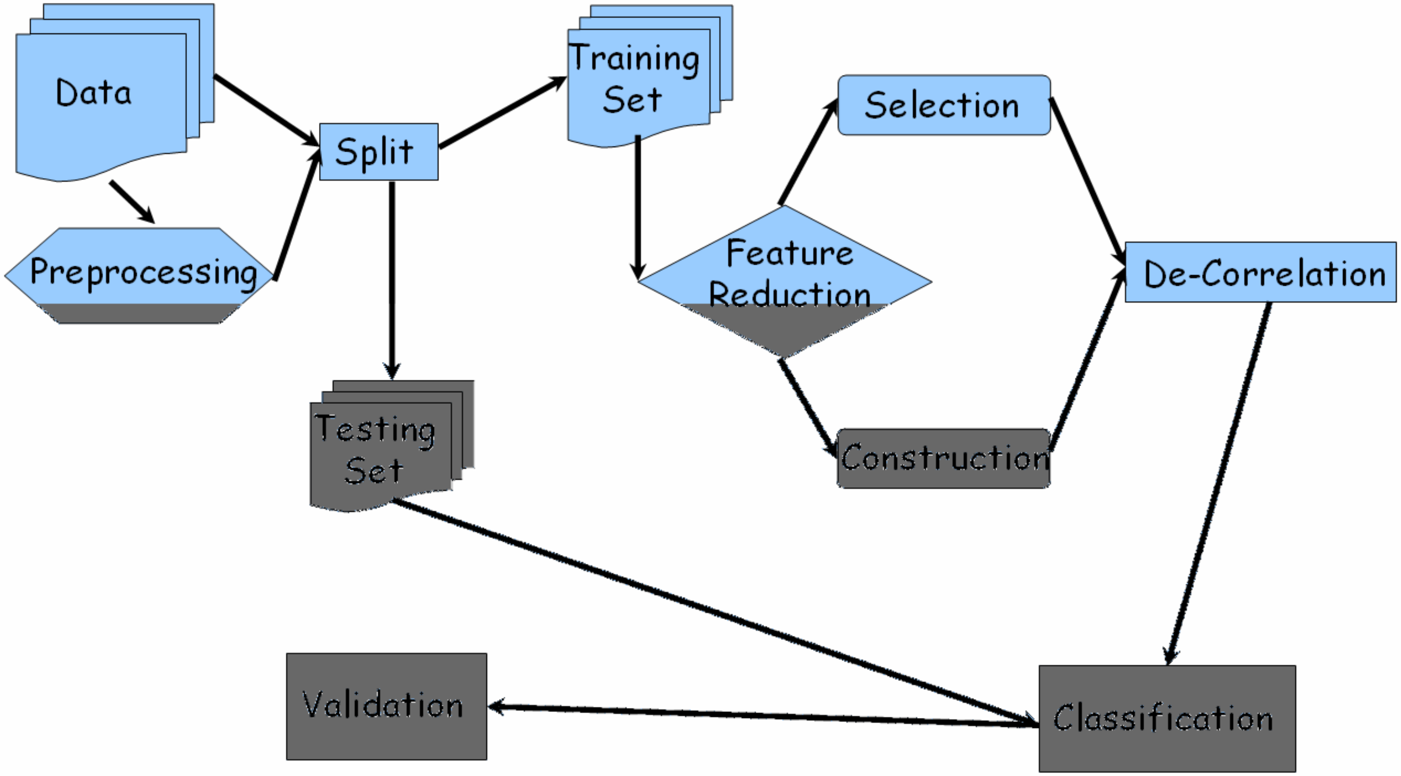




Special Capabilities

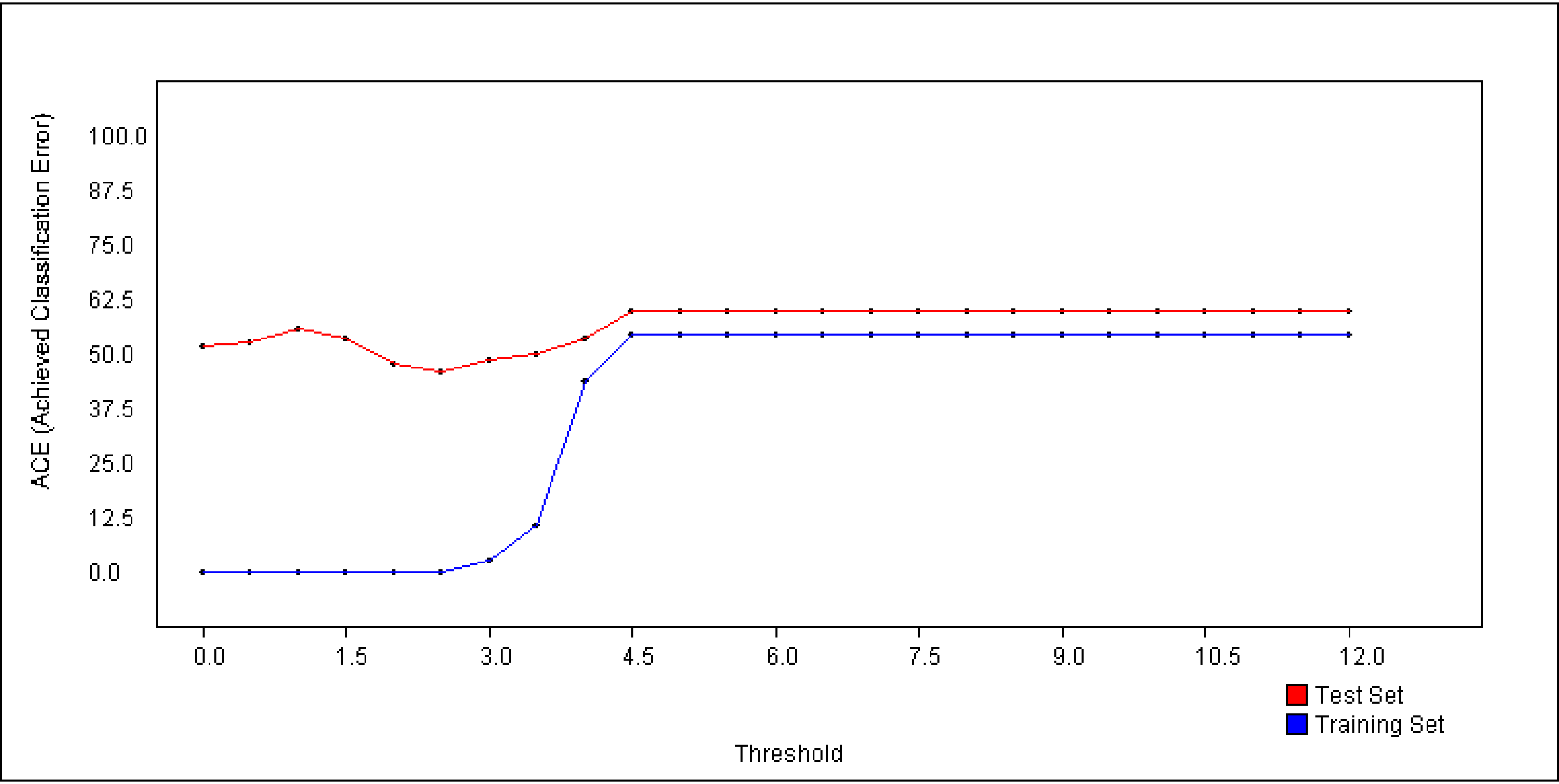
- Built to facilitate comparisons of methods of analysis via cross-validation + other methods
- Computation validation methods include:
  - Nonparametric bootstrapping
  - Leave-one-out validation
  - Random Resampling Validation
  - *k*-fold validation (to be added)
  - *Efficiency Analysis*** NEW*
- Gene Expression Pattern Grid
- Proof-by-Pubmed *on the fly*

Framework of Evaluation



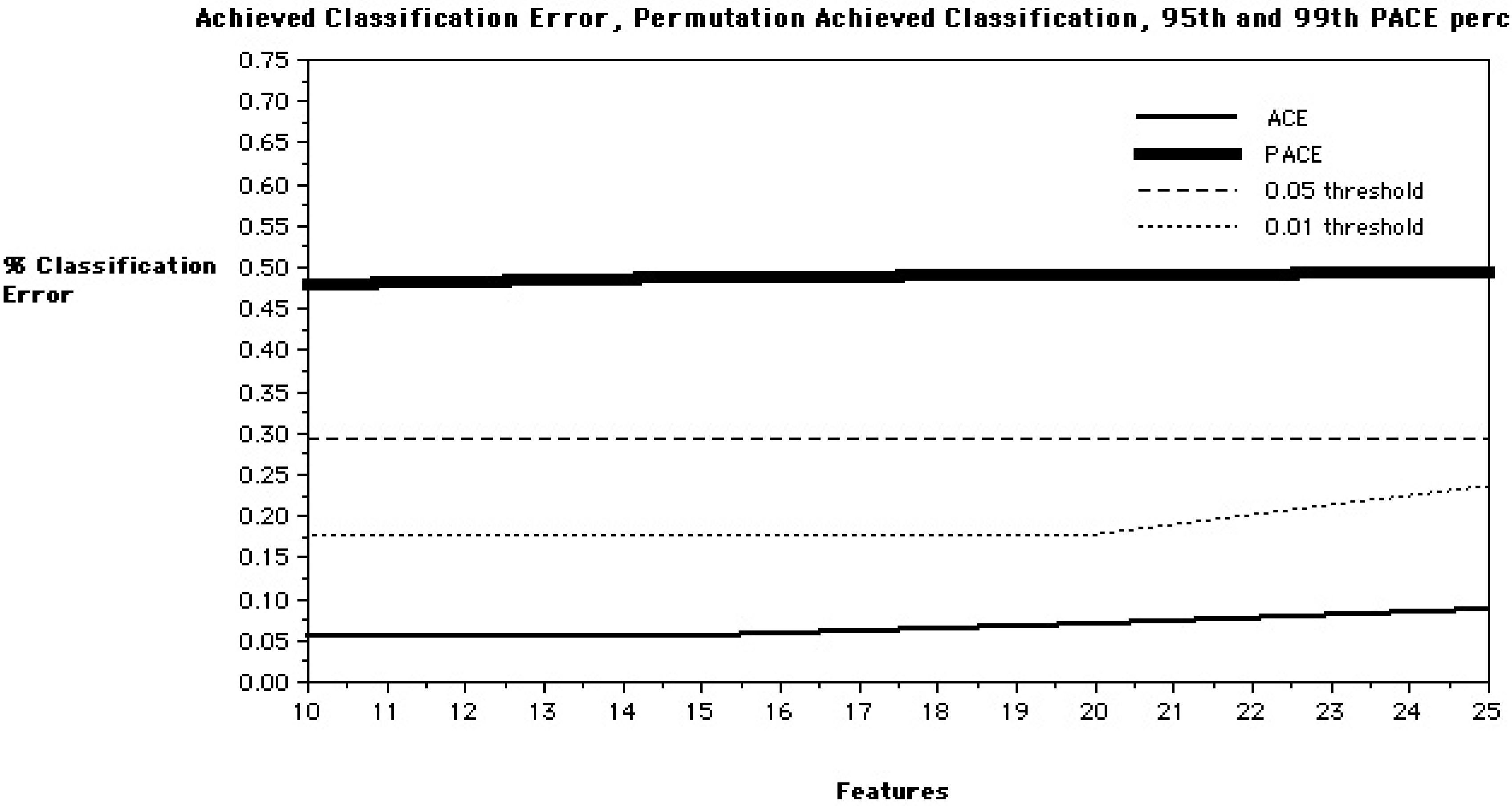
Credit: Richard Pelikan

Random Resampling

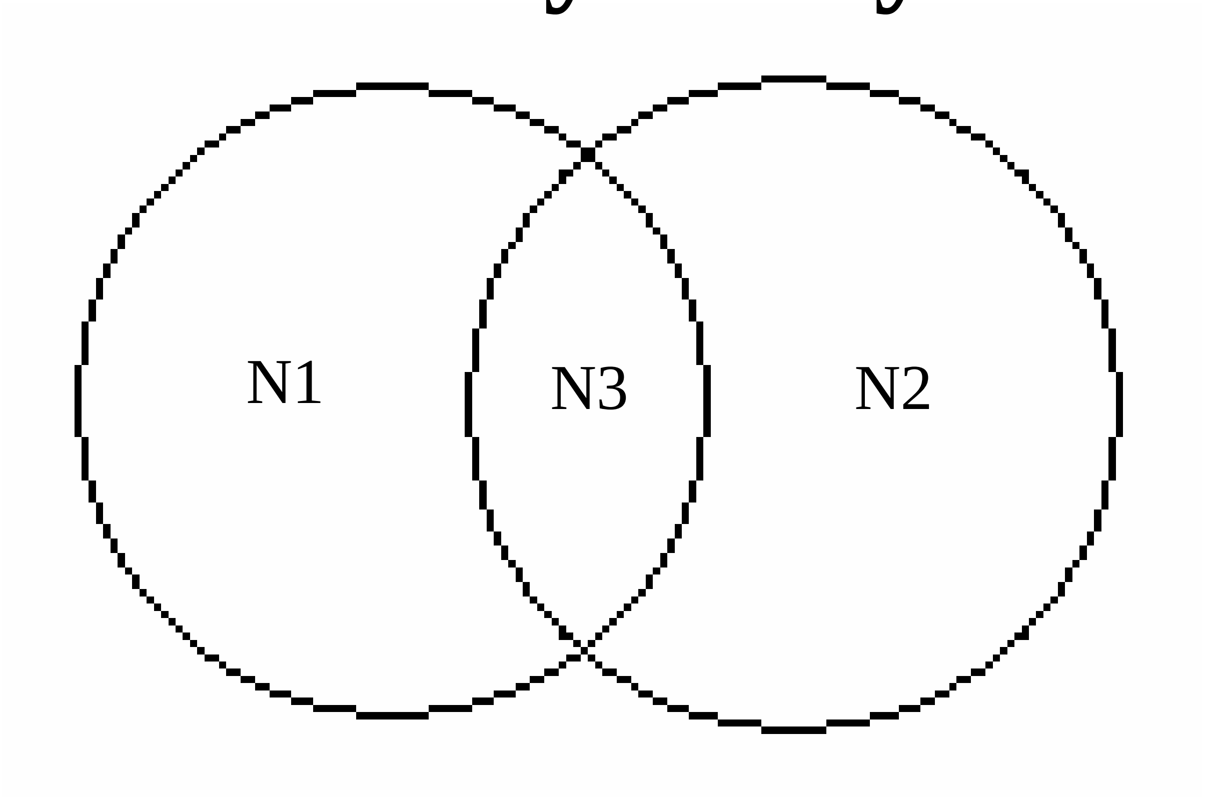


Random data set N1 = N2 = 16; 1100 random ‘genes’; t-test

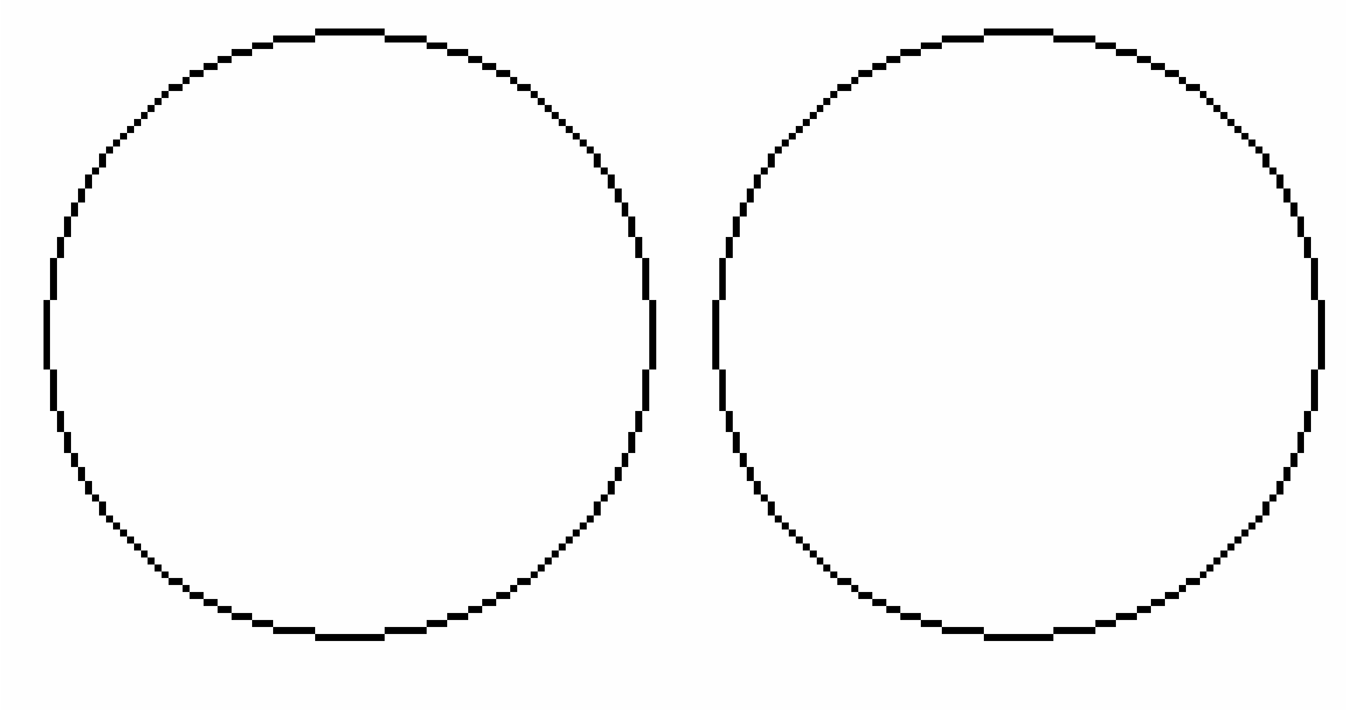
Significance of achieved classification error



Efficiency Analysis

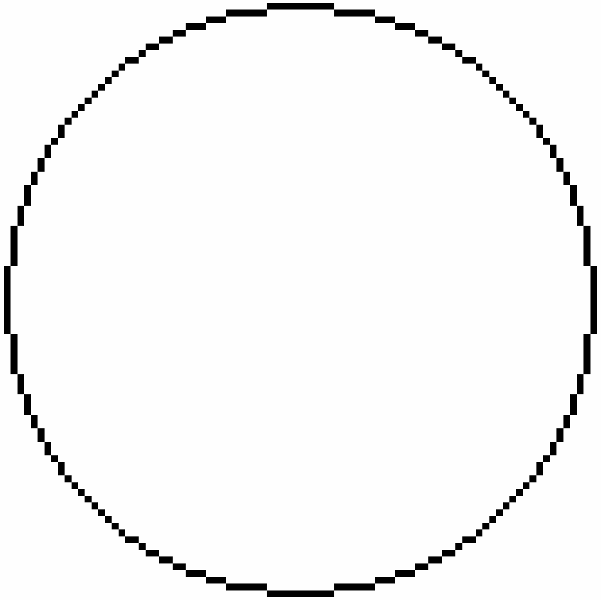


O = (2*N3)/(N1+N2)



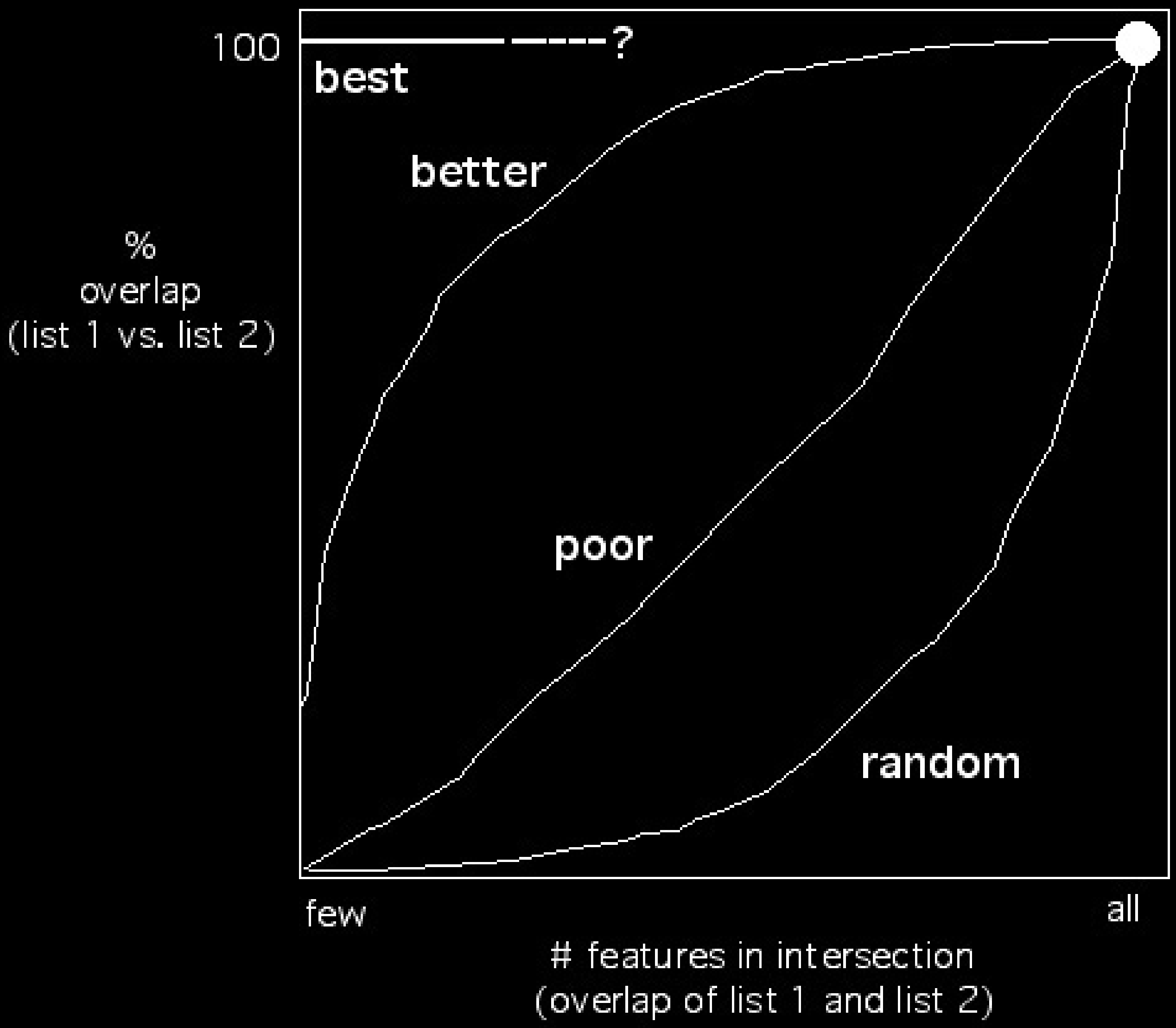
O = (2*N3)/(N1+N2)

N3 = 0; O = 0



O = (2*N3)/(N1+N2)

N3 = N1+N2; O = 1.0


100
best
?
better
poor
random
%
overlap
(list 1 vs. list 2)
few
all
# features in intersection
(overlap of list 1 and list 2)

Astrocytoma Progression Markers
Early Stage vs. Late Stage

- **USA**
- Khatua et al.
- Early: N = 7
- Late: N = 8
- Genes: 8497
- Journal: *Cancer Res.*

[CANCER RESEARCH 63, 1865–1870, April 15, 2003]

Overexpression of the *EGFR/FKBP12/HIF-2α* Pathway Identified in Childhood Astrocytomas by Angiogenesis Gene Profiling[1,2]

Soumen Khatua,[3] Katia M. Peterson,[3] Kevin M. Brown,[3] Christopher Lawlor, Maria R. Santi, Bonnie LaFleur, Devin Dressman, Dietrich A. Stephan, and Tobey J. MacDonald[4]

Center for Cancer Research, Children's Research Institute [S. K., K. M. P., C. L., T. J. M.]; Research Center for Genetic Medicine, Children's Research Institute [K. M. B., D. D., D. A. S.], and Department of Pathology [M. R. S.], Children's National Medical Center, Washington, DC 20010; Department of Preventive Medicine, Division of Biostatistics, Vanderbilt University, Nashville, Tennessee 37232 [B. L.]; and Graduate Program in Genetics, George Washington University, Washington, DC 20052 [K. M. B.]

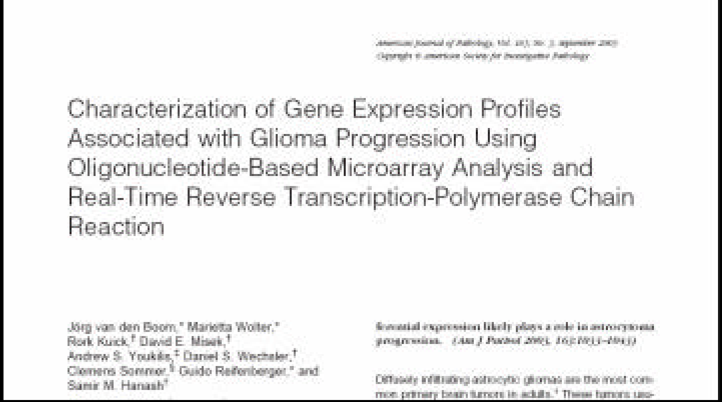
ABSTRACT

Intense angiogenic proliferation, a histopathological hallmark distinguishing malignant from benign astrocytomas, is vital for tumor progression. Thus, identifying and targeting specific pathways that promote malignant astrocytoma-induced angiogenesis could have substantial therapeutic benefit. Expression profiling of 13 childhood astrocytomas to determine the expression pattern of 133 angiogenesis-related genes revealed that 44 (33%) genes were differentially expressed (17 were overexpressed, and 27 were underexpressed) between malignant high-grade astrocytomas (HGAs) and benign low-grade astrocytomas. Hierarchical clustering and principal components analysis using only the 133 angiogenesis-related genes distinguished HGA from low-grade astrocytomas in 100% of the samples analyzed, as did unsupervised analysis using the entire set of 8198 expressed genes represented on the array, indicating that the angiogenesis-related genes were reliable markers of pathological

in which overall survival remains less than 30% (2). Thus, novel therapeutic approaches are needed for childhood HGA.

Studies demonstrating the crucial role of angiogenesis in cancer have been a major advance in our understanding of malignant tumor progression (3). One of the key histopathological features that distinguish HGA from LGA is intense, increased angiogenesis. The invasiveness of HGA, another unique feature of this tumor in comparison with LGA, is associated with increased microvascular density and intratumoral hypoxia (4). Thus, inhibitors of hypoxia-inducible angiogenesis factors could be important new therapeutic agents against HGA. In adults, the most commonly described regulators of brain tumor-derived angiogenesis are VEGF, platelet-derived growth factor, angiopoietin-2, and their respective receptors (5). It is not known to what extent these same regulatory mechanisms exist in pediatric

- **Germany**
- van den Boom et al.
- Early: N = 8
- Late: N = 8
- Genes: 5682
- Journal: *Am J Pathol.*



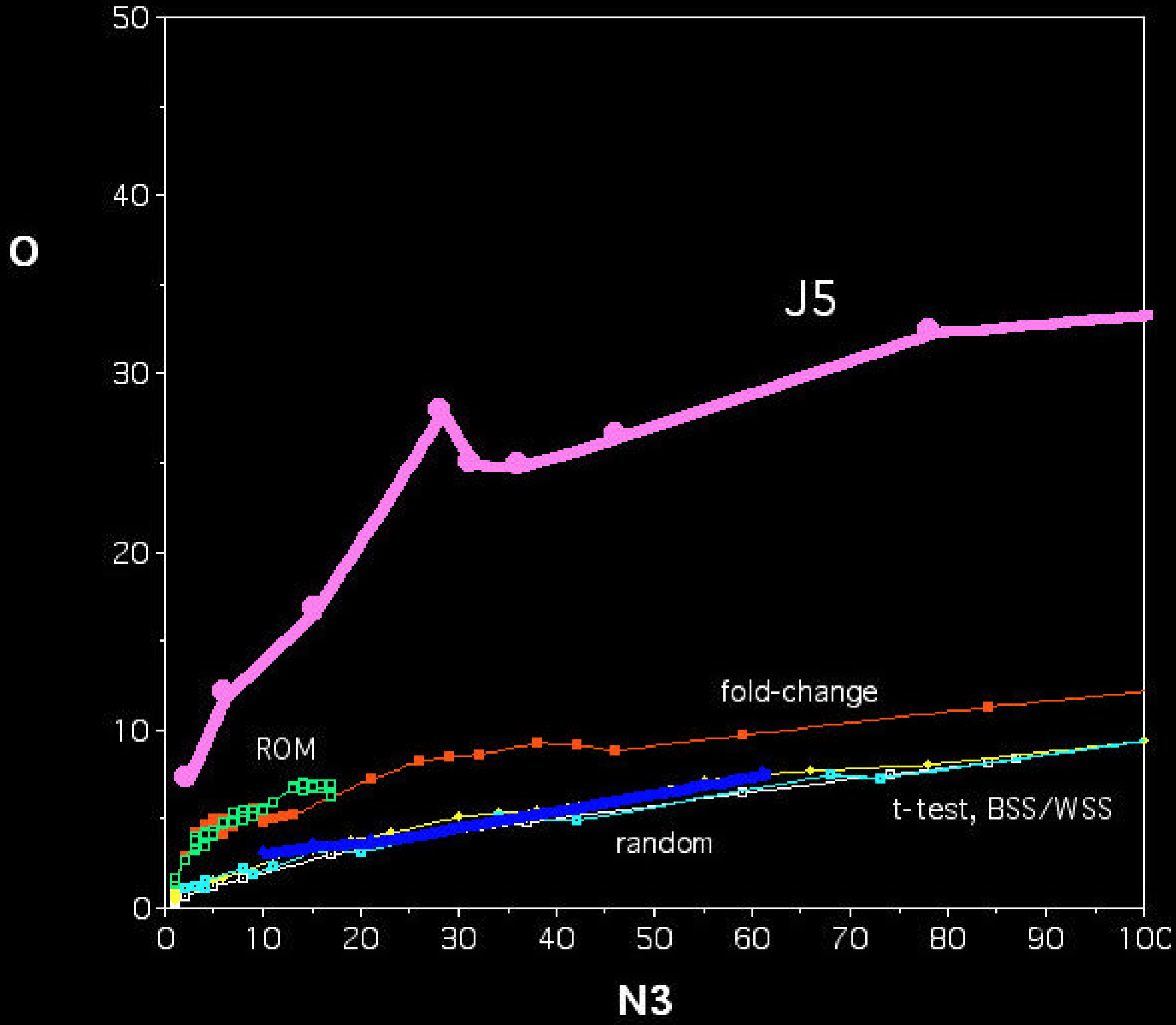
American Journal of Pathology, Vol. 163, No. 3, September 2003
Copyright © American Society for Investigative Pathology

Characterization of Gene Expression Profiles Associated with Glioma Progression Using Oligonucleotide-Based Microarray Analysis and Real-Time Reverse Transcription-Polymerase Chain Reaction

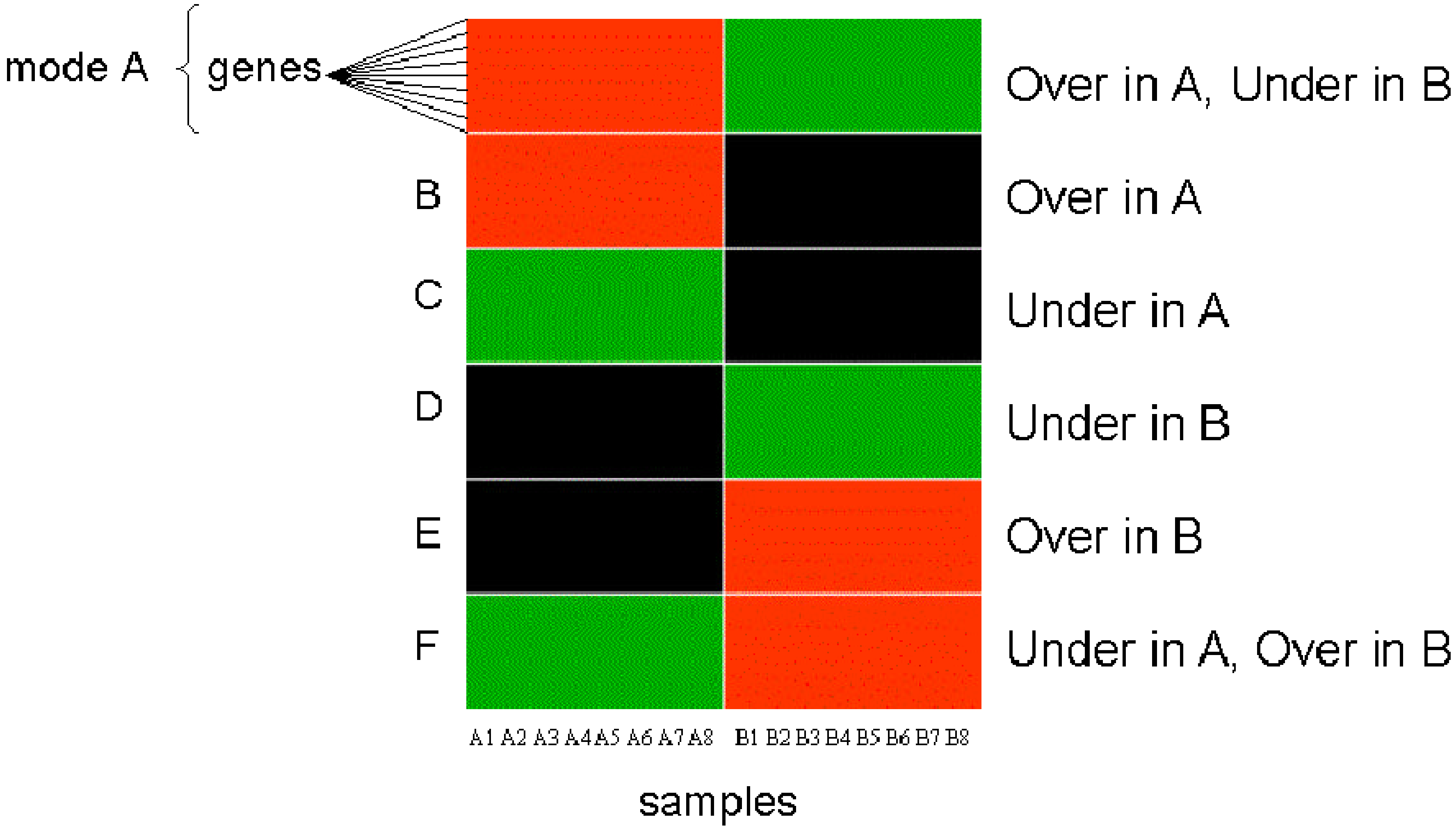
Jörg van den Boom,* Marietta Wolter,* Rork Kuick,† David E. Misek,† Andrew S. Youkilis,‡ Daniel S. Wechsler,† Clemens Sommer,§ Guido Reifenberger,* and Samir M. Hanash†

ferential expression likely plays a role in astrocytoma progression. *(Am J Pathol 2003, 163:1033–1043)*

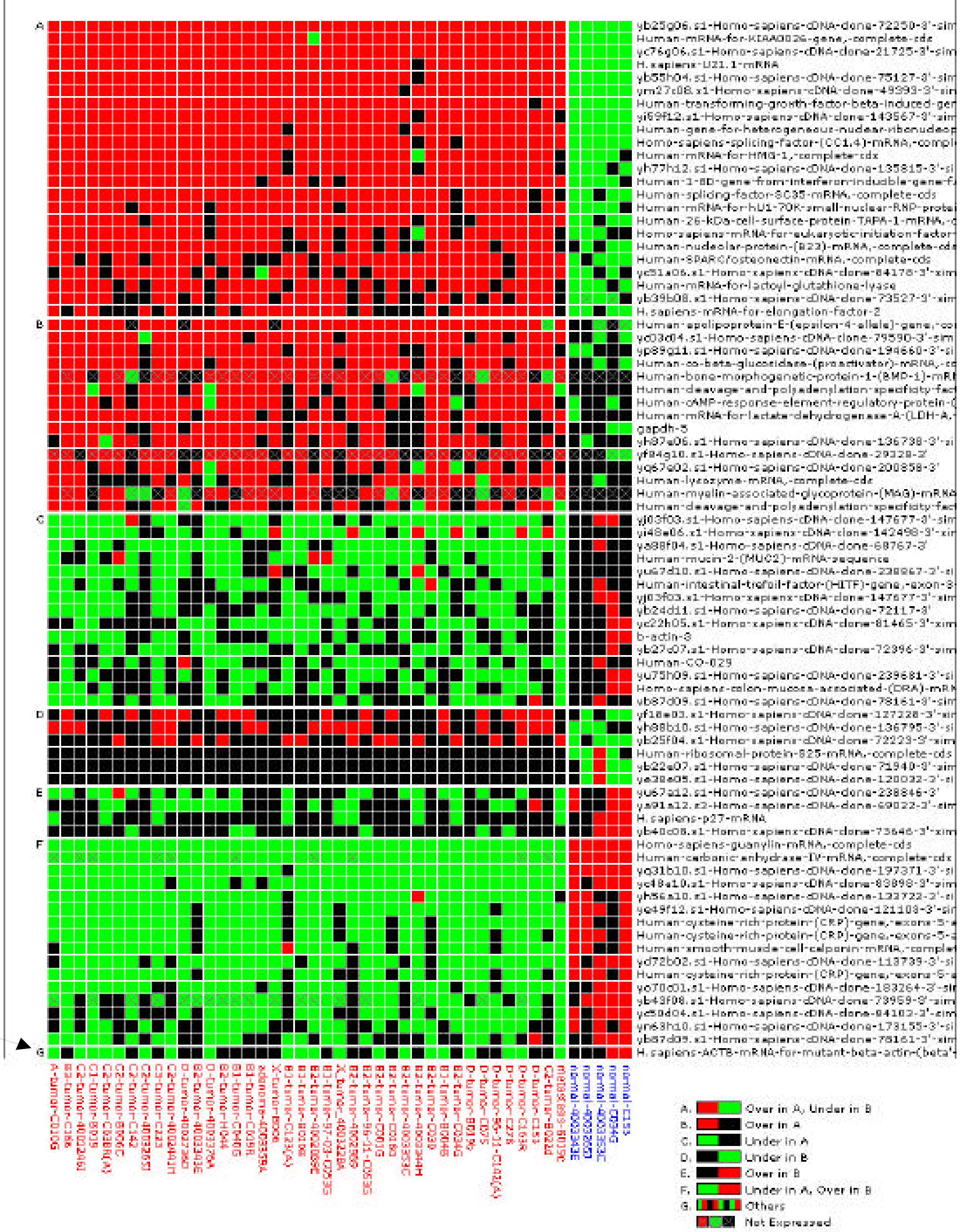
Diffusely infiltrating astrocytic gliomas are the most common primary brain tumors in adults.[1] These tumors usu-


50
40
30
20
10
0
O
J5
fold-change
ROM
t-test, BSS/WSS
random
0
10
20
30
40
50
60
70
80
90
10C
N3

Gene Expression Pattern Grid



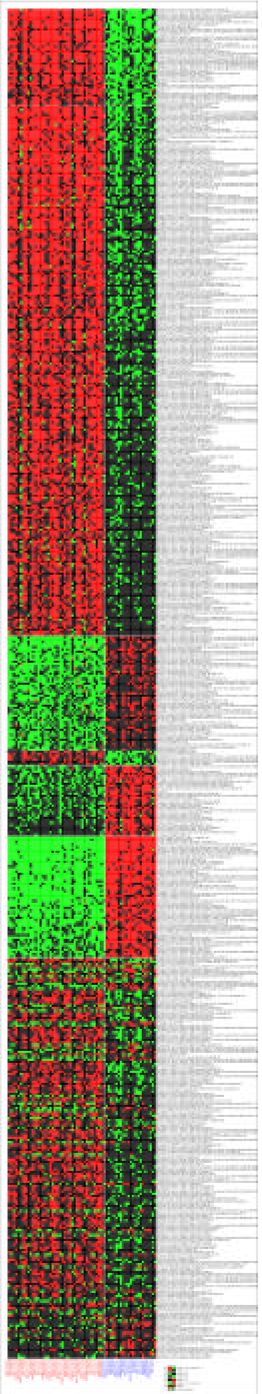
‘G’ = other


A
B
C
D
E
F
G
yb25g06.s1-Homo-sapiens-cDNA-clone-72250-3'-sim
Human-mRNA-for-KIAA0026-gene,-complete-cds
yc76g06.s1-Homo-sapiens-cDNA-clone-21725-3'-sim
H.sapiens-U21.1-mRNA
yb55h04.s1-Homo-sapiens-cDNA-clone-75127-3'-sim
ym27c08.s1-Homo-sapiens-cDNA-clone-49393-3'-sim
Human-transforming-growth-factor-beta-induced-gen
yi69f12.s1-Homo-sapiens-cDNA-clone-143567-3'-sim
Human-gene-for-heterogeneous-nuclear-ribonucleop
Homo-sapiens-splicing-factor-(CC1.4)-mRNA,-compl
Human-mRNA-for-HMG-1,-complete-cds
yh77h12.s1-Homo-sapiens-cDNA-clone-135815-3'-si
Human-1-8D-gene-from-interferon-inducible-gene-f
Human-splicing-factor-SC35-mRNA,-complete-cds
Human-mRNA-for-hU1-70K-small-nuclear-RNP-protei
Human-26-kDa-cell-surface-protein-TAPA-1-mRNA,-c
Homo-sapiens-mRNA-for-eukaryotic-initiation-factor
Human-nucleolar-protein-(B23)-mRNA,-complete-cds
Human-SPARC/osteonectin-mRNA,-complete-cds
yc51a06.s1-Homo-sapiens-cDNA-clone-84178-3'-sim
Human-mRNA-for-lactoyl-glutathione-lyase
yb39b08.s1-Homo-sapiens-cDNA-clone-73527-3'-sim
H.sapiens-mRNA-for-elongation-factor-2
Human-apolipoprotein-E-(epsilon-4-allele)-gene,-co
yc03d04.s1-Homo-sapiens-cDNA-clone-79590-3'-sim
yp89g11.s1-Homo-sapiens-cDNA-clone-194660-3'-si
Human-co-beta-glucosidase-(proactivator)-mRNA,-co
Human-bone-morphogenetic-protein-1-(BMP-1)-mRN
Human-cleavage-and-polyadenylation-specificity-fact
Human-cAMP-response-element-regulatory-protein-(
Human-mRNA-for-lactate-dehydrogenase-A-(LDH-A,-
gapdh-5
yh87e06.s1-Homo-sapiens-cDNA-clone-136738-3'-si
yf84g10.s1-Homo-sapiens-cDNA-clone-29328-3'
yq67e02.s1-Homo-sapiens-cDNA-clone-200858-3'
Human-lysozyme-mRNA,-complete-cds
Human-myelin-associated-glycoprotein-(MAG)-mRNA
Human-cleavage-and-polyadenylation-specificity-fact
yj03f03.s1-Homo-sapiens-cDNA-clone-147677-3'-sim
yi48e06.s1-Homo-sapiens-cDNA-clone-142498-3'-sim
ya88f04.s1-Homo-sapiens-cDNA-clone-68767-3'
Human-mucin-2-(MUC2)-mRNA-sequence
yu67d10.s1-Homo-sapiens-cDNA-clone-238967-3'-si
Human-intestinal-trefoil-factor-(HITF)-gene,-exon-3
yj03f03.s1-Homo-sapiens-cDNA-clone-147677-3'-sim
yb24d11.s1-Homo-sapiens-cDNA-clone-72117-3'
yc22h05.s1-Homo-sapiens-cDNA-clone-81465-3'-sim
b-actin-3
yb27c07.s1-Homo-sapiens-cDNA-clone-72396-3'-sim
Human-CO-029
yu75h09.s1-Homo-sapiens-cDNA-clone-239681-3'-si
Homo-sapiens-colon-mucosa-associated-(DRA)-mRN
yb87d09.s1-Homo-sapiens-cDNA-clone-78161-3'-sim
yf18e03.s1-Homo-sapiens-cDNA-clone-127228-3'-sim
yh88b10.s1-Homo-sapiens-cDNA-clone-136795-3'-si
yb25f04.s1-Homo-sapiens-cDNA-clone-72223-3'-sim
Human-ribosomal-protein-S25-mRNA,-complete-cds
yb22e07.s1-Homo-sapiens-cDNA-clone-71940-3'-sim
ye38e05.s1-Homo-sapiens-cDNA-clone-120032-3'-si
yu67a12.s1-Homo-sapiens-cDNA-clone-238846-3'
ya91a12.s2-Homo-sapiens-cDNA-clone-69022-3'-sim
H.sapiens-p27-mRNA
yb40c08.s1-Homo-sapiens-cDNA-clone-73646-3'-sim
Homo-sapiens-guanylin-mRNA,-complete-cds
Human-carbonic-anhydrase-IV-mRNA,-complete-cds
yq31b10.s1-Homo-sapiens-cDNA-clone-197371-3'-si
yc48a10.s1-Homo-sapiens-cDNA-clone-83898-3'-sim
yh56a10.s1-Homo-sapiens-cDNA-clone-133722-3'-si
ye49f12.s1-Homo-sapiens-cDNA-clone-121103-3'-sim
Human-cysteine-rich-protein-(CRP)-gene,-exons-5-a
Human-cysteine-rich-protein-(CRP)-gene,-exons-5-a
Human-smooth-muscle-cell-calponin-mRNA,-complet
yd72b02.s1-Homo-sapiens-cDNA-clone-113739-3'-si
Human-cysteine-rich-protein-(CRP)-gene,-exons-5-a
yo70c01.s1-Homo-sapiens-cDNA-clone-183264-3'-sim
yb43f08.s1-Homo-sapiens-cDNA-clone-73959-3'-sim
yc50d04.s1-Homo-sapiens-cDNA-clone-84103-3'-sim
yn63h10.s1-Homo-sapiens-cDNA-clone-173155-3'-si
yb87d09.s1-Homo-sapiens-cDNA-clone-78161-3'-sim
H.sapiens-ACTB-mRNA-for-mutant-beta-actin-(beta'-
A-tumor-C010G
B3-tumor-C186
C2-tumor-4002246J
C1-tumor-B019
C2-tumor-C188R(A)
C2-tumor-B606C
C2-tumor-C142
C2-tumor-4002365J
C3-tumor-C223
C2-tumor-4002441J
D-tumor-4002728D
B2-tumor-4003341E
D-tumor-4003370A
B2-tumor-H044
B1-tumor-C140G
B1-tumor-C019A
adenoma-4003359A
X-tumor-B006
B3-tumor-C123(A)
B3-tumor-B018B
B2-tumor-4002086E
B3-tumor-97-03-C0536
X_tumor_4003228A
B2-tumor-4002509
B2-tumor-96-11-C0535
B3-tumor-C001G
B2-tumor-C018D
B2-tumor-4003353C
B3-tumor-4003344H
B2-tumor-C030
B3-tumor-B004S
B2-tumor-C034G
D-tumor-B019b
D-tumor-C075
D-tumor-96-11-C102(A)
D-tumor-C276
D-tumor-C163R
D-tumor-C151
C2-tumor-B024d
metastasis-B015C
normal-4003343E
normal-4003260D
normal-4003353C
normal-C034G
normal-C153
A. Over in A, Under in B
B. Over in A
C. Under in A
D. Under in B
E. Over in B
F. Under in A, Over in B
G. Others
Not Expressed

Some tests lead to more sizeable 'G' group which, while statistically significant, exhibit no coherent signs of differential expression in most samples.

Outliers or conflicting patterns of differential expression.

(colon cancer data set, t-test, cut-point = 4.0)



Priorities

- **Enhance!**
- **Integrate and Interoperate!**
- **Annotate!**
- **Blow it up!**
- **Characterize and represent**
  - Data models
  - Schema
  - UML Diagrams:
    - Use case diagrams
    - Activity diagrams
    - Sequence diagrams
    - Package diagrams…

Priorities

- **Enhance!**
  - Increase data format diversity tolerance
  - Add outlier spot detection, adopt existing QC criteria
  - Add normalization (e.g., DWD), tests, classification methods
  - Apply Jprogram (Duke) to allow assimilation of R projects
  - Add pathway analysis and interaction analysis capabilities (cMAP, cPATH, cytoScape…)

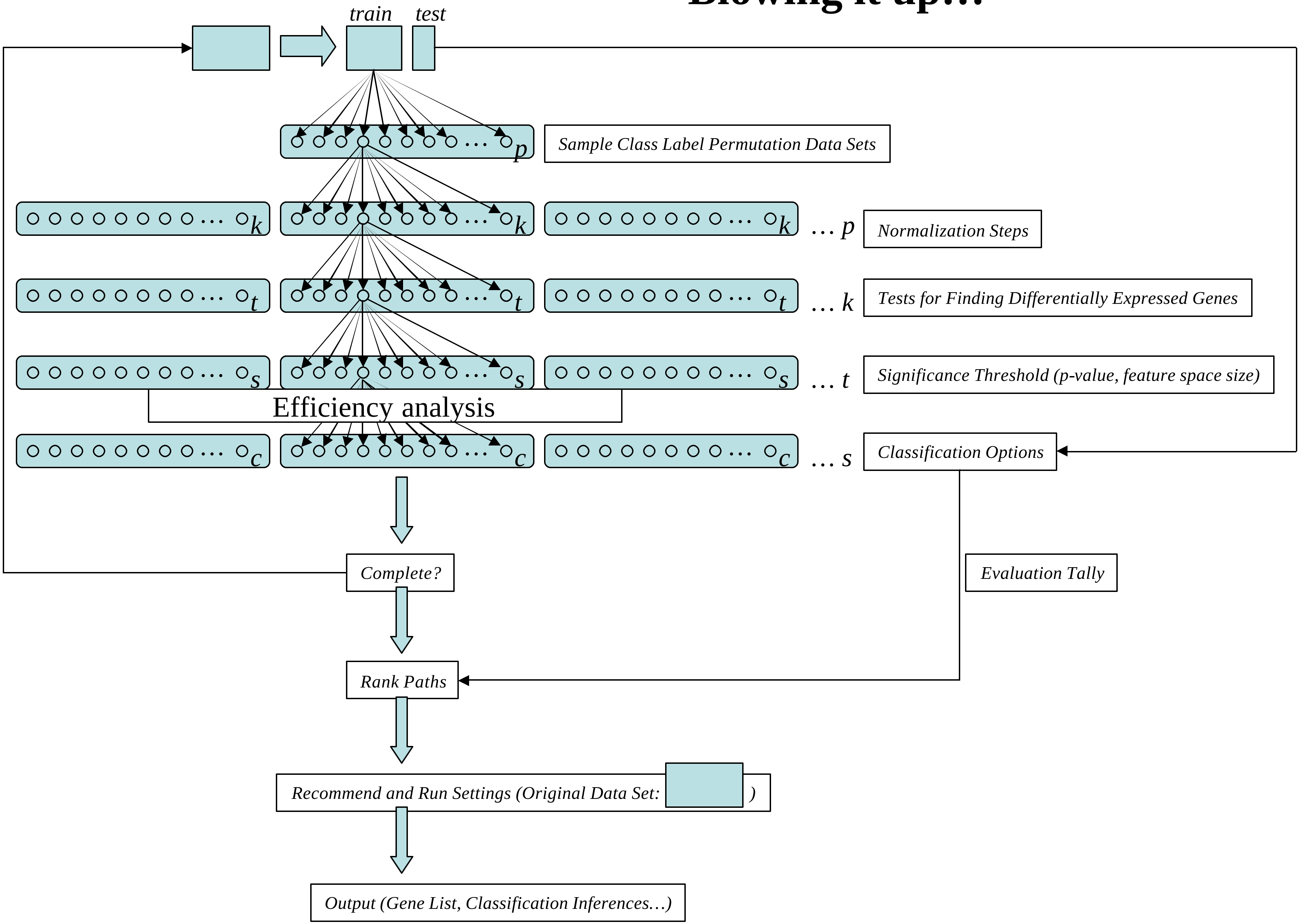
Priorities

- **Integrate and Interoperate!**
  - SPOT, SPROC, LIMS projects, OncoMine, FDGP could produce data dumps in caGEDA formats - or adopt html interface that finds an active caGEDA server (local or on the grid) for on-the-fly analysis
  - caGEDA could output in formats or make direct calls to:
    - GoMiner
    - cPATH
    - GKB (Reactome project)

Priorities

- **Annotate!**
  - – Five components:
    - English text description
    - Mathematical description
    - Pseudocode
    - Source code
    - Related literature

Blowing it up…



Sample Class Label Permutation Data Sets

Normalization Steps

Tests for Finding Differentially Expressed Genes

Significance Threshold (p-value, feature space size)

Efficiency analysis

Classification Options

Complete?

Evaluation Tally

Rank Paths

Recommend and Run Settings (Original Data Set:)

Output (Gene List, Classification Inferences…)

Demo

- **http://bioinformatics.upmc.edu/GE2/GEDA.html**

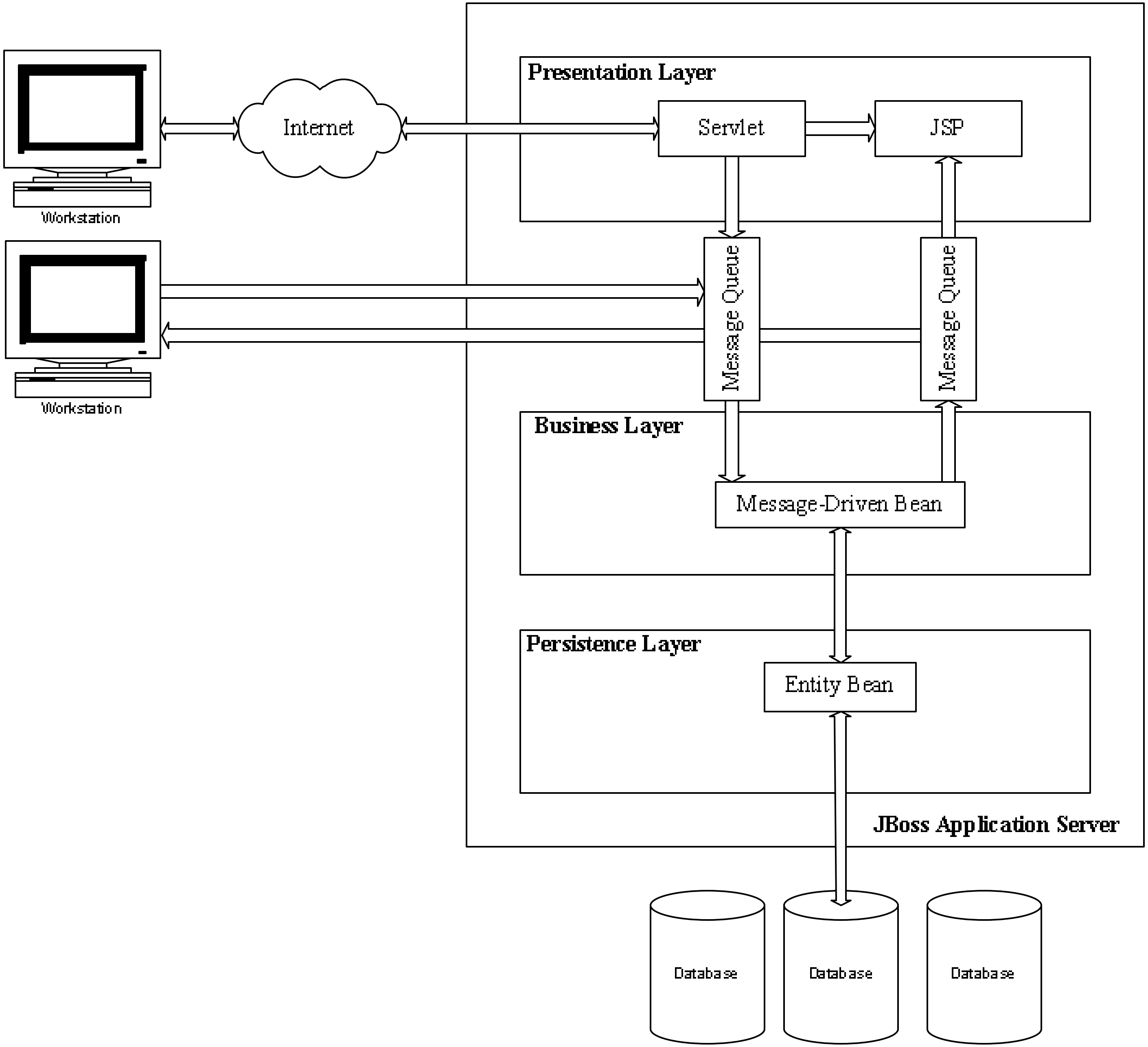
Development (credit: S. Patel)

caGEDA application development is an iterative software development approach that leverages elements from Rational Unified Process (RUP). Use cases for the application are developed using the expertise available at the research center. Once the use case analysis is completed, an iterative functional design and development process is applied, which allows for rapid and segmented development of the application. During the iteration, all the software development activities are executed. The artifacts associated with each functional iteration includes: detailed use cases describing the function; class and sequence diagrams; a system architecture diagram; the actual software code; a project plan describing subsequent iterations; and a test plan for software validation.

- UML modeling and use case development is performed using UML modeling tool from Rational Rose. Source code is developed using the Java programming libraries for Servlet, JSP and EJB. We use Apache software's Ant to assist the software build process. All the server side software components are tested on JBoss application server. All the software components used in development the GEDA application are freely available on the Internet.

Application Architecture

caGEDA conforms to n-tier architectural design that include several layers. A presentation layer includes a web application server that transforms the request coming from the Internet browser in to the calls to the business logic and provides programmatic access to the application. A Business layer can communicate to the standalone application client directly using RMI-IIOP protocol or using CORBA. Presentation layer objects communicate with the business layer objects using RMI-IIOP protocol. Since the communication with the business layer can be done using CORBA even a non-java application client can make use of the services provided by caGEDA.


Workstation
Workstation
Internet
Presentation Layer
Servlet
JSP
Message Queue
Message Queue
Business Layer
Message-Driven Bean
Persistence Layer
Entity Bean
JBoss Application Server
Database
Database
Database

Presentation tier involves one or more web servers, each responsible for interacting with end user. The presentation tier displays the requested information in HTML to the end user; it also reads and interprets the user's selection and makes invocations to the business tier's components. The implementation of presentation tier uses Servlets and JSPs

Business tier consists of multiple EJB components running under the hood of EJB container/server. These are reusable components that are independent of any user interface logic. We should be able to, for example, take our business tier and port it to different presentation tier (such as application client) with no modification. Our business tier is made up of session, entity and message-driven beans.

Data tier is where the permanent data resides. With use of entity beans, we can leave our options to use virtually any database of choice. Switching to database of a particular choice should be seamlessly achievable.

Software Life Cycle: Iterative approach

High level requirement analysis
Scope
Data Format
Data Pre -processing
Feature Selection
Prediction
Computational Validation
Data Visualization
Databases
Security
Architecture and Design
Estimation & Schedule
Iterative Design & Development
Testing
Continuous testing and integration
Coding standards
Implementation
Testing
Beta Testing
Feedback
Deployment
Deployment plan
User documentation
Bug reporting/tracking system

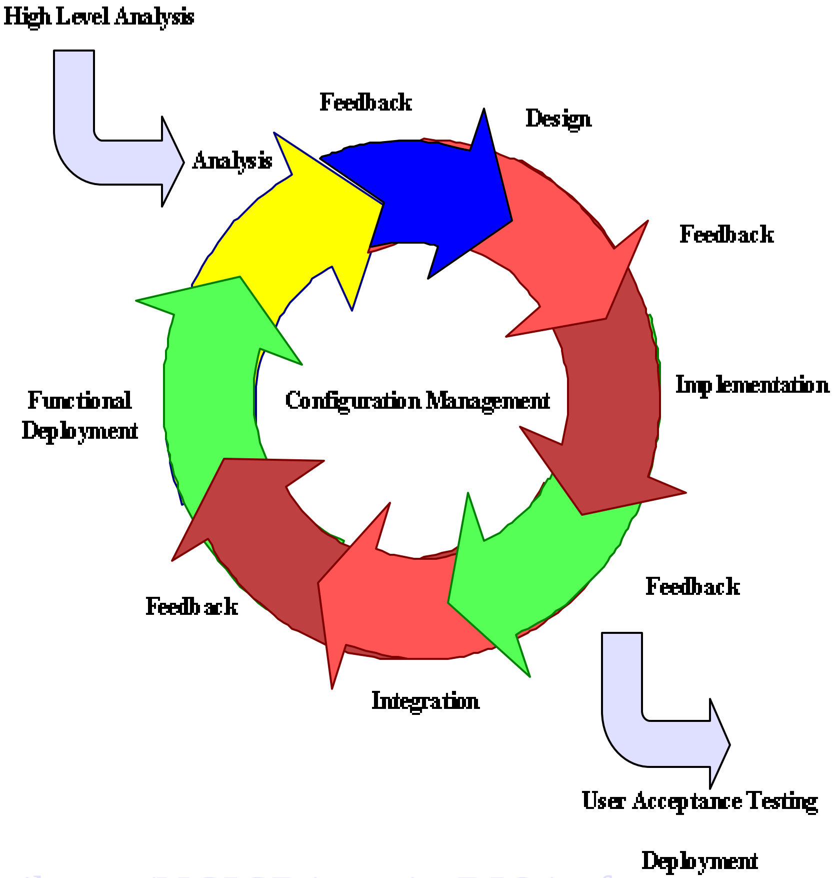


Figure source: http://ncicb.nci.nih.gov/NCICB/core/caBIO/software_process